Is the Polarity of Content Producers Strongly Influenced by the Results of the Event?

Liliana Ibeth Barbosa Santillán
University of Guadalajara
Information Technology
Guadalajara, México
Email: ibarbosa@cucea.udg.mx

Inmaculada Álvarez de Mon y Rego
Universida Politécnica de Madrid
Lingüistica aplicada a la ciencia y la tecnología
Madrid, Spain
Email: ialvarez@euitt.upm.es

***Abstract*—This paper presents an approach to compare two types of data, subjective data (Polarity of Pan American Games 2011 event by country) and objective data (the number of medals won by each participating country), based on the Pearson correlation. When dealing with events described by people, knowledge acquisition is difficult because their structure is heterogeneous and subjective. A first step towards knowing the polarity of the information provided by people consists in automatically classifying the posts into clusters according to their polarity. The authors carried out a set of experiments using a corpus that consists of 5600 posts extracted from 168 Internet resources related to a specific event: the 2011 Pan American games. The approach is based on four components: a crawler, a filter, a synthesizer and a polarity analyzer. The PanAmerican approach automatically classifies the polarity of the event into clusters with the following results: 588 positive, 336 neutral, and 76 negative. Our work found out that the polarity of the content produced was strongly influenced by the results of the event with a correlation of .74. Thus, it is possible to conclude that the polarity of content is strongly affected by the results of the event. Finally, the accuracy of the PanAmerican approach is: .87, .90, and .80 according to the precision of the three classes of polarity evaluated.**

***Keywords*—Polarity; Subjective; Objective Corpus Analysis.**

I. INTRODUCTION

Content producers are currently emerging from the social web where the majority of the population is young people with very specific needs in terms of communication. The Web has facilitated social networking phenomena through both structured and unstructured data. The analysis of this content may have considerable influence on important decisions that affect society.

The amount of Internet resources that exist in the Web dealing with a specific event, such as newspapers, chat rooms, social networking, Internet commerce, product reviews and blogs, have heterogeneous content that proliferates in an uncontrolled fashion.

However, there are difficulties in measuring the polarity of the content generated by a person. Some of them are: a) identifying noise polarity and fake reviews and b) dealing with slang and inaccurate use of language.

In addition, one of the most important and complex tasks for the entrepreneurs, officials and the organizers of an event is to know as precisely as possible how citizens perceive it. In this sense, the diversity of opinions and assessments related to events that involve different countries vary greatly. A possible solution is to obtain metrics for measuring the polarity of the content expressed by producers in their writings.

The motivation of this research is to develop a first approach on how to assess the appreciation of an event through the opinions of citizens grouped by their origin country for an event of Pan-American scale.

This study focuses on answering the following research question: Is the polarity of content producers strongly influenced by the results of the event?

The hypothesis that the authors propose is:

H1 The polarity of content producers is strongly influenced by the results of the event.

This study aims to provide benefit in the form of classify the social point of view of polarity of the 2011 Pan American Games.

This project could be of benefit for both governments and citizens.

Compared with previous work, the major contributions of this paper are the following:

- Focusing on unstructured opinions in order to contrast subjective polarity and objective data related to the same event.

This work draws on at least 5,600 opinions of citizens of the Pan American countries, helping to understand the impact of the event among the citizens of the 42 nations participating in 36 sports. Some of the advantages of our analysis are that it facilitates knowing the polarity of the citizens through fresh opinions articulated in Internet resources.

This paper is structured as follows. Section II briefly discusses the related work. Section III gives an overview of the design, describing the proposed PanAmerican approach. Sections IV and V discuss the analysis and experiments. Finally, Section VI contains the conclusions of our research work.

II. RELATED WORK

The related work takes into account two topics: 2.1) polarity, and 2.2) systems related to olympics.

 ISBN: 978-1-61208-311-7

A. Polarity

According to Cambria [1], Opinion Mining "mainly concerns polarity detection", whereas sentiment analysis, as defined by Pang [2], is "the specific application of classifying reviews as to their polarity (either positive or negative)". Opinion mining and sentiment analysis are used in this research as synonyms in order to deal with the literature related to both topics.

In recent years, opinion mining has been studied by many researchers. The authors have focused on three aspects: a) question answering, b) recommendation systems, and c) sentiment-relevant lexicons.

- **Question answering:** Earlier work showed that disambiguating instances of subjectivity clues is useful for sentence-level attitude-type classification. Somasundaran et al. [3] developed automatic classifiers to recognize when a sentence is expressing one of the two main types of attitude. Stoyanov [4] developed a corpus of opinion questions and answers; his research compared and contrasted the properties of facts and opinions in question answering. Vlad et al. [5] defined qualitative dimensions for evaluating answers and showed how ignored terms in the process of entity definition can help users to discover underlying information.
- **Recommendation systems:** Efforts in this area were carried out by Nitin et al. [6] who proposed a collaborative exploration system helping users to explore movie reviews from various viewpoints. Reputation is a topic of collective interest; Morinaga [7] demonstrated that it is possible to help users to discover important knowledge regarding the reputations of products of interest through the following tasks: characteristic word extraction, co-occurring word extraction, sentence extraction, and correspondence analysis. Ungar et al. [8] used clustering methods for collaborative filtering.
- **Sentiment-relevant lexicons:** Previous research has focused on the creation of lexicons in English such as that of: Higashinaka [9] who used a set of dialogues to build her own lexicon. Lexicons are also available as linguistic resources on the Internet, some examples being: SentiWordnet [10], NTU Sentiment Dictionary [11]. Pak [12] build automatically sentiment relevant lexicon from Internet Resources, and the OpinionFinder system for subjectivity analysis [13], among others.

B. Systems

The research of Gruzd et al. [14] measures if happiness is contagious online in 2010 winter olympics and they determined that were more positive messages than negative in twitters. It also influenced the level of retweet from positive versus negative messages. SentiStrength [15] splits the tweets into positive and negative conversations and filters them through a programme, which systematically converts them into a lightshow. It was used for measure the Olympic London Eye.

As opposed to these works, we aim to model multiple users' location posts and learn polarity from numerous opinions by different individuals on the Pan American Games 2011.

III. The PanAmerican approach Proposed

The PanAmerican approach aims at performing the classification of polarity in a set of Internet resources focused on an event. The approach is based on four components: a crawler (A), a filter (B), a synthesizer (C), and a polarity analyzer (D). The main function of the crawler component is to search and find data from internet resources related to the event of interest. After locating the data, the filter component processes the data in order to remove noise. The filter component only debugs internet resources that are associated with the event. At this point, the corpus consists of numerous posts containing large amounts of data from many countries and in many languages. The synthesizer component represents the amount of data into clusters with similar expressions using unsupervised learning. Finally, the Polarity analyzer component classifies each cluster into positive, neutral or negative. Each of the components in the PanAmerican approach is described in greater detail below:

A. Crawler

The crawler is a component that obtains internet resources related to one event and stores them in a repository for later use.

The main challenges faced by this component are the size of the internet resources that continue to grow in a highly dynamic way and the fact that some of them appear and disappear in a very brief period of time. The depth is the link levels that a seed can have; if a seed has another link embedded in the body of the web this will conduct a search on this link, and so on up to the number of levels configured on the system. It is important to note that an unlimited number of levels can take months of processing in multiple threads. The solution is to seek and obtain Internet resources based on a depth of four links [16].

In addition, the crawler component stores items in a knowledge base as follows: the URL, the contents of the internet resource in natural language, a bag of words and the position in the sentence of each one of these.

An important challenge is the quality of the internet resources because over 40 percent of the data collected is not semantically related to the event under study. One of the reasons is that many people manipulate their content specifically with keywords, titles, and descriptions in order for their Internet resource to rank high in search results during the event and for that reason the filter component is essential.

1) Data: The quality of the corpus is measured by the degree of compliance of the posts that meet the purpose for which the corpus is compiled. Thus it was necessary to take special care in the selection of posts attempting to maintain homogeneity. Therefore, it was necessary to establish the following criteria that govern the selection and inclusion of posts.

Quantity: It was decided to include 5600 posts of different dimensions.

 ISBN: 978-1-61208-311-7

Quality of text: Given that the selection was automatic, special care was taken in that the texts were written in the correct language, without spelling mistakes, in clear writing.
Published in Pan American Games 2011: Due to the nature of the project, we only included published posts.
Type of opinion: the opinion must have been carried out with the results of the event.
Text form: The texts must be written in the form of general impressions.
Style: The texts must be comprehensive, describing the opinion from beginning to end, discarding free or incomplete texts introduced in unfinished or abandoned posts.
Additional information: Each sample must be marked with a series of additional data, which gives extra information and allows for identification. These marks are the: web page from which it has been extracted, country or area where the opinion has been realized, language, and date of the opinion.

B. Filter

The filter is a component that processes Internet resources to remove unwanted data related to an event. The filter uses an anti-noise function to minimize 37.5 percent of the noise in the corpus content. The filter works by analyzing the most frequent words in the post titles and descriptions that are related to the name of the event and then classifies each post as noise or not noise. Finally, the filter builds a knowledge base populated with titles, descriptions and posts related to the event. The output from the filter component is still a large volume of data because there are one hundred sixty-eight Internet resources that are producing dozens of posts daily in several languages. As a result, the knowledge base grows tremendously and is full of instances that were saved sequentially; thus, the next step is a preprocessing of all the resulting data, grouping and classifying them according to common themes using the synthesizer component.

C. Synthesizer

The main function of the synthesizer component is to take all the posts that the filter has classified as not noise and without previous knowledge about them identify groups of similar expressions using a Bayes classifier [17].

Therefore, the next step is to group all the posts that express similar content to represent the data in a lower dimension space. One reason to use unsupervised learning [18] is that people observe the same event in several ways; however, their perceptions of one event have clear differences based on their nationality. We used a Bayes classifier, which has shown good results in previous work. The results are stored in two plain text files: one with a set of clusters, and the other one with the six most representative patterns for each cluster of posts. The synthesizer component creates a new population of posts represented in a lower dimension space by clustering.

D. Polarity Analyzer

The polarity analyzer component is in charge of the polarity analysis of clusters based on a Multilingual Lexical Ontology (MLO) (see section 1) and classifies each cluster into positive, negative or neutral. It uses K-means cluster.

For each cluster, we obtained the six most representative patterns: the component performs a semantic analysis of patterns based on the MLO ontology.

Therefore, subjectivity is calculated with an additional operation as follows: positive (more than zero), negative (less than zero), and neutral (equal to zero).

Finally, each cluster is classified into positive, negative or neutral.

1) Multilingual Lexical Ontology (MLO): The two main characteristics of the MLO ontology are that it is language-independent and provides multilingual population in any language. However, their instances are in the four languages, these being Spanish, English, Portuguese and French because they are the languages most used by people in the 2011 Pan American Games.

Definition: the MLO Ontology is a conceptual description based on a lexicon of the subjective words in Natural Language as shown in (1). The MLO Ontology consists of four disjoint sets C, R, A, and τ where C means concept identifiers (2), R means relation identifiers (3 and 4), A means attribute identifiers (5), and τ means data types (6).

$$MLO := (C, \leq c, R, \gamma_R, \leq_R, A, \gamma_A, \tau) \tag{1}$$

The set C of concepts is:

$$C := \left\{ \begin{array}{l} Adjectives, NegativeAdjectives, \\ PositiveAdjectives, Adverbs, NegativeAdverbs \\ , PositiveAdverbs, Articles, Authors, \\ DomainResources, Nouns, NegativeNouns, \\ PositiveNouns, Paragraphs, Posts, \\ Predicates, Prepositions, Sentences, \\ Subjects, Titles, InternetResources, \\ Verbs, NegativeVerbs, PositiveVerbs \end{array} \right. \tag{2}$$

The set R of relations is:

$$R := \left\{ \begin{array}{l} author\_of, post\_of, paragraph\_of, sentence\_of, \\ adverb\_in, articles\_inprepositions\_in, nouns\_in, \\ adjectives\_in, verbs\_in, subject\_of, predicate\_of \end{array} \right. \tag{3}$$

where the relation hierarchy defines that DomainResources has the relation author_of that belongs to Authors. InternetResources has the relation post_of that belongs to Posts, following the same logic the rest of the relations are defined, as shown in equation (4).

$$\begin{array}{l} \gamma R(author\_of) = (Authors, DomainResources) \\ \gamma R(post\_of) = (Posts, InternetResources) \\ \gamma R(paragraph\_of) = (Paragraphs, Posts) \\ \gamma R(sentence\_of) = (Sentences, Paragraphs) \\ \gamma R(adverbs\_in) = (Adverbs, Sentences) \\ \gamma R(articles\_in) = (Articles, Sentences) \\ \gamma R(prepositions\_in) = (Prepositions, Sentences) \\ \gamma R(nouns\_in) = (Nouns, Sentences) \\ \gamma R(adjectives\_in) = (Adjectives, Sentences) \\ \gamma R(verbs\_in) = (Verbs, Sentences) \\ \gamma R(subject\_in) = (Subjects, Sentences) \\ \gamma R(predicate\_in) = (Predicates, Sentences) \end{array} \tag{4}$$

The set A of attribute identifiers is:

$$A := \left\{ \begin{array}{l} blog, author, title, post, paragraph, sentence, \\ subject, predicate, article, noun, nounP, nounN, \\ verb, verbN, verbP, adjective, adjectiveP, \\ adjectiveN, preposition, adverb, adverbP, \\ adverbN \end{array} \right. \tag{5}$$

The set τ of datatypes contains only one element a string, as shown in (6).

$$\tau := (string) \tag{6}$$

The first axiom defines the concept NegativeAdverbs as equivalent to saying that there is a negativeAdverb, which stands in a adverb_in relation with the corresponding sentence, following the same logic the rest of the axioms are defined as shown in (7).

$$\begin{aligned}
&\forall x(NegativeAdverbs(x) \longleftrightarrow \exists y \wedge adverb_in(x,y) \wedge Sentences(y))\\
&\forall x(PositiveAdverbs(x) \longleftrightarrow \exists y \wedge adverb_in(x,y) \wedge Sentences(y))\\
&\forall x(NegativeVerbs(x) \longleftrightarrow \exists y \wedge verbs_in(x,y) \wedge Sentences(y))\\
&\forall x(PositiveVerbs(x) \longleftrightarrow \exists y \wedge verbs_in(x,y) \wedge Sentences(y))\\
&\forall x(NegativeNouns(x) \longleftrightarrow \exists y \wedge nouns_in(x,y) \wedge Sentences(y))\\
&\forall x(PositiveNouns(x) \longleftrightarrow \exists y \wedge nouns_in(x,y) \wedge Sentences(y))\\
&\forall x(NegativeAdjectives(x) \longleftrightarrow \exists y \wedge adjectives_in(x,y) \wedge Sentences(y))\\
&\forall x(PositiveAdjectives(x) \longleftrightarrow \exists y \wedge adjectives_in(x,y) \wedge Sentences(y))
\end{aligned} \tag{7}$$

To summarize, the PanAmerican approach proposed is shown in Fig. 1, where the input is the Official Web of Pan American Games 2011 and the output is the polarity value of each country involved.

```
1:  procedure CRAWLER(SubsetWeb)
2:      for i ← 1, SizeSubsetWeb do
3:          InternetResources(i) ← DownloadURL((Get(URL(i)));
4:      end for
5:  end procedure
6:  procedure FILTER(InternetResources,Term)
7:      for i ← 1, NumberofInternetResources do
8:          if SyntacticFilter(InternetResource(i)) then
9:              if NoiseFilter(InternetResource(i)) then
10:                 noise(InternetResource(i)) ← InternetResource((i));
11:             else
12:                 if SemanticFilter(InternetResource(i)) then
13:                     Titles(InternetResource(i)) ← Split(Titles(InternetResource(i)));
14:                     Descriptions(InternetResource(i)) ← Split(Descriptions(InternetResource(i)));
15:                     Posts(InternetResource(i)) ← Split(Posts(InternetResource(i)));
16:                 end if
17:             end if
18:         end if
19:     end for
20: end procedure
21: procedure SYNTESIZER(Posts)
22:     for i ← 1, NumberofPosts do
23:         ClusterMultilingual(i) ← ClusterMultilingual(Posts(i));
24:         Patterns(i) ← Patterns(Posts(i));
25:     end for
26: end procedure
27: procedure POLARITY(Web)
28:     InternetResources ← Crawler(Web);
29:     Posts ← Filter(InternetResources,"PanAmericangames2011");
30:     Clusters ← Syntesizer(Posts);
31:     for k ← 1, NumberofClusters do
32:         PolarityValue(k) ← PolarityValue(Cluster(k), MLO);
33:         Sum(k) ← OpinionValue(k);
34:     end for
35: end procedure
```

Fig. 1: The PanAmerican approach

IV. EXPERIMENTAL DETAILS AND PERFORMANCE RESULTS

The following section includes a detailed description of how the experiment was conducted. The first part describes the objectives of the experiment and the second part focuses on the results obtained after the experiment was conducted.

The experimental setup had four objectives: 1) to obtain a subset of Internet resources related to the 2011 Pan American Games, 2) to delete noise in the Internet resources obtained, 3) to classify sets of posts that are close in meaning and group them into clusters, and 4) to assess the polarity of each cluster.

The analysis was carried out in two ways: 1) the crawler component was run in order to obtain Internet resources over a period of three months, and 2) the type of PanAmerican was carried out for 168 Internet resources.

The first task was to obtain posts in Internet resources using the crawler component. However, the output is a set of Internet resources that grows rapidly with data irrelevant to the analysis. For that reason it was necessary to adjust the crawler component to the event. The focus was placed on Internet resources related syntactically to the term "Pan American games 2011" in order to reject those Internet resources, which were not related to this event. We deal with two problems: a) the seed [19] had 300 related Internet resources so the crawler component was restricted to a search of four levels deep and b) identifying the source country based on the meaning of its posts is a major task. Thus, we assumed that, depending on the Internet resources, the top level domain of the post was used for the country of the author.

The second task was to filter out Internet resources not related semantically to the event and as a result classified as noise by the filter component. For example, some of the following Internet resources had a Pan American 2011 term but not all contained data related to the event; there are noise traders who use the same term but with a different meaning:

```
{http://www.emailbrain.com/134087/rss,
http://feeds2.feedburner.com/noc-aho\,
http://www.argentina.ar/rss/rss_prensa_es.xml,
http://www.dushi-curacao.info/1-dushi-curacao.html,
http://www.bahamasolympiccommittee.org/_rss/news,
http://www.amandala.com.bz/inc/rss.php?id=11806,
http://bmxbolivia.org/?feed=rss2,
http://www.olympic.ca/fr/,
http://co.elpais.feedsportal.com/c/33807/f/607321/index.rss,
http://juegospanamericanos.ain.cu/feed/,
http://www.colimdo.org/rss.aspx,
http://ministryofhealth.gov.ky/feed/rss.xml,
http://www.elcaribe.com.do/rss,
http://www.avn.info.ve/rss/6 , etc. }
```

As a result, we obtained a corpus of 3500 posts extracted from one hundred sixty-eight different Internet resources, sampled from a comprehensive range of 2011 Pan American Games Internet resources with 147 MB of text.

The third task was to find group similarities so as to represent in a lower dimension space the PanAmerican analysis. The PanAmerican approach synthesizes the corpus into clusters with similar data. As an example, we show three posts that are grouped to their similar themes:

```
Commonwealth Youth Games Team Selected,
Team Bahamas Deports Mexico with 3 Medals,
BOC Ammounces Guadalajara 2011 Pan Am Games Team
```

The PanAmerican approach processes each one of the clusters and extracts the six most relevant patterns for the task of PanAmerican; these patterns are also assessed through a parser that performs semantic matching between each pattern and the MLO ontology in order to carried out the tagging of polarity.

 ISBN: 978-1-61208-311-7

The MLO ontology measures are shown in Table I where the numbers of local instances are the same for all the four languages involved. For example, PositiveVerbs in Spanish amount to one hundred instances and this number is the same for each of the other languages: English, Portuguese and French.

TABLE I: MULTILINGUAL LEXICAL ONTOLOGY (MLO) MEASURES

Type	Measures	
	Local	*Inferred*
$NumberofTriples$	11616	1
$NegativeAdjectives$	1092	0
$PositiveAdjectives$	1468	0
$NegativeAdverbs$	52	0
$PositiveAdverbs$	100	0
$Internetresources$	286	0
$NegativeNouns$	880	0
$PositiveNouns$	800	0
$Titles$	1000	0
$DomainInternetResources$	1	0
$NegativeVerbs$	472	0
$PositiveVerbs$	400	0
$owl:Class$	22	0
$owl:DatatypeProperty$	22	0
$owl:NamedIndividual$	5400	0
$owl:ObjectProperty$	13	0
$owl:Ontology$	1	0

Each cluster is classified using a well-known formula of the sum of the value of the patterns polarity that is shown in (8).

$$f(n) = \begin{cases} n > 0 & \text{if } n \text{ is Positive (P)} \\ n < 0 & \text{if } n \text{ is Negative (N)} \\ n = 0 & \text{if } n \text{ is Neutral (Z)} \end{cases} \quad (8)$$

A partial result is shown in Table II where Cluster 2 (C2) contains posts, which are linked to people in wheelchairs, despite the fact that the first post is not explicitly linked to that term. However, Tito Bautista was a participant with a wheelchair, so it is correctly specified. The C2 PanAmerican result is positive (P) polarity.

The C6 language is French and most of the posts are negative; therefore, the PanAmerican analyzer component value is negative (N) polarity.

In C5, the first two posts are in Portuguese and the third post is in Spanish; all of these are linked to the Brasil term. The first two posts are neutral polarity and the third is negative polarity and as a result the cluster is negative polarity.

TABLE II: A PARTIAL VIEW OF CLUSTERS COMPONENT OUTPUT

Clusters	Cluster of Posts
C1 (P)	México llega a la Villa Miranda de México, una favorita de Guadalajara Respalda el Presidente Calderón propuesta de EGM para buscar los Juegos Olímpicos para Jalisco
C2 (P)	Perseverancia y coraje, palabras que definen a Tito Bautista Seleccionados mexicanos liderearon el Circuito Nacional de Tenis en Silla de Ruedas El Tenis en Silla de Ruedas entrará en acción
C3 (Z)	Registration [closed] Sport Management Course Start of CAC Games 2010 Pan American Games 2011
C4 (P)	Team Bahamas Departs Mexico with 3 Medals DIF Jalisco y COPAG hacen mancuerna Commonwealth Youth Games Team Selected
C5 (N)	Lettre de démission du Ministre de la Justice Affaire Bélizaire : Rapport de la Commission Spéciale dú2019Enqueate (Partie 1) Brasil equipo a vencer en Voleibol Sentados
C6 (N)	Mission afghane: départ devancé de lúAustralie? NY: Un homme aurait voulu faire sauter des sites Tripoli veut juger Seif al-Islam en Libye

A. Performance Results

In the first place, the crawler component identified four Internet resources for each participating country in the 2011 Pan American Games and therefore obtained 168 Internet resources containing 5600 posts as shown in Fig. 2

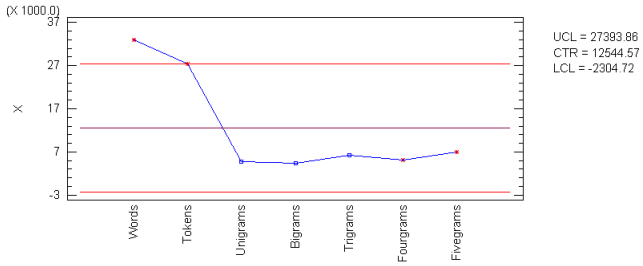

Fig. 2: Identification of the 5600 posts structure based on number of words, tokens, unigrams, bigrams, trigrams, fougrams, and fivegrams.

At this point, the filter component was applied in order to delete posts, which contained noise (2100) and posts semantically related to the event (3500) were identified. Next, 1000 clusters were obtained using the synthesizer component. Finally, 588 positive clusters, 336 neutral clusters and 76 negative clusters were tagged using the polarity analyzer component.

To measure the accuracy of the cluster classification task we used well-known formulae in the area of information retrieval as shown in equations 9 through 12, where precision and recall were evaluated for each polarity (P, Z, N).

Precision was calculated by dividing the True Positives (TP) between the sum of True Positives and False Positives (FP) as shown in (11). Recall is the division between True Positives and the sum of True Positives and False Positives as shown in (9).

 ISBN: 978-1-61208-311-7

$$Recall \equiv TPRate = TP/(TP + FN)) \quad (9)$$

$$FPRate = FP/(FP + TN)) \quad (10)$$

$$Precision = TP/(TP + FP)) \quad (11)$$

$$F - Measure = 2TP/(2TP + FP + FN)) \quad (12)$$

The results for precision for each polarity -Positive, Negative and Neutral (P, N, Z) respectively- are shown in Table III. The highest accuracy is in the negative polarity with a value of .9.

TABLE III: DETAILED ACCURANCY OF POLARITY COMPONENT

TP Rate	FP Rate	Precision	Recall	F-Measure	Polarity
0.997	0.204	0.875	0.997	0.932	P
0.118	0.001	0.9	0.118	0.209	N
0.762	0.096	0.8	0.762	0.78	Z

The clusters that were correctly classified amount to 85.1% and the faulty were 14.9% . The sample comprised 1000 clusters derived from 3500 posts that were filtered from 168 Internet resources from 42 countries and 4 languages. The absolute and relative errors are also shown in Table IV.

TABLE IV: STRATIFIED CROSS-VALIDATION OF COMPONENT SYNTHESIZER

Correctly Classified Clusters	85.1 %
Incorrectly Classified Clusters	14.9 %
Kappa statistic	0.7007
Mean absolute error	0.1881
Root mean squared error	0.2929
Relative absolute error	52.6328 %
Root relative squared error	69.3158 %
Total Number of Clusters	1000

The PanAmerican results and the medals won for each country are shown graphically in Fig. 3a and, as it can be seen, the positive clusters dominate. Fig. 3b shows the polarity results for each country.

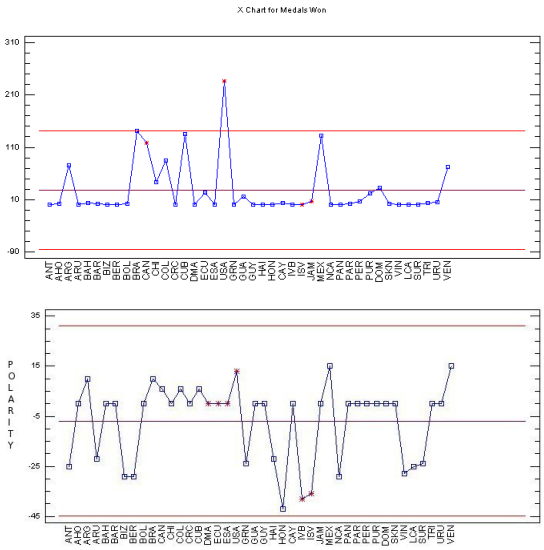

Fig. 3: Medals won in Pan American 2011 Games and Polarity Results of PanAmerican Approach for each country.

where the Id for each country is in Table V.

TABLE V: ID FOR EACH COUNTRY

Name	Id	Name	Id
Antigua and Barbuda	(ANT)	Guyana	(GUY)
Netherlands Antilles	(AHO)	Haiti	(HAI)
Argentina	(ARG)	Honduras	(HON)
Aruba	(ARU)	Cayman Islands	(CAY)
Bahamas	(BAH)	Virgin Islands (GB)	(IVB)
Barbados	(BAR)	Virgin Islands(US)	(ISV)
Belize	(BER)	Jamaica	(JAM)
Bermudas	(ANT)	Mexico	(MEX)
Bolivia	(BOL)	Nicaragua	(NCA)
Brazil	(BRA)	Panama	(PAN)
Canada	(CAN)	Paraguay	(PAR)
Chile	(CHI)	Peru	(PER)
Colombia	(COL)	Puerto Rico	(PUR)
CostaRica	(CRC)	Dominican Republic	(DOM)
Cuba	(CUB)	Saint Kitts Nevis	(SKN)
Dominica	(DMA)	Saint Vincent and the Grenadines	(VIN)
Ecuador	(ECU)	Saint Lucia	(LCA)
El Salvador	(ESA)	Suriname	(SUR)
United States of America	(USA)	Trinidad and Tobago	(TRI)
Grenada	(GRN)	Uruguay	(URU)
Guatemala	(GUA)	Venezuela	(VEN)

In addition, the research hypothesis claimed that the assessments of content producers would be influenced strongly by the results of an event regardless of their nationality. From our results it can be seen that the appraisal in some countries was positive because of the high number of medals won, as in the case of the United States, which took 236 medals. This is in contrast with Honduras, which did not win in any field, and thus, the overall assessment was strongly influenced and negatively evaluated, such as is shown in Fig. 3b. Fig. 4 shows that the correlation coefficient is equal to 0.74, indicating a strong relationship between the medals won and the global polarity by each country.

 ISBN: 978-1-61208-311-7

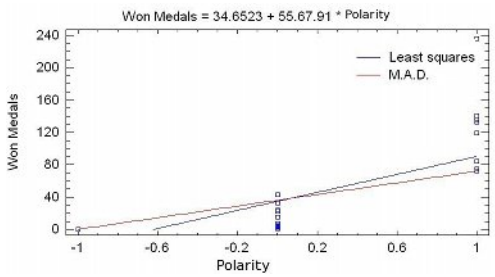

Fig. 4: Shows that the correlation coefficient is equal to 0.74, indicating a strong relationship between the medals won in Pan American 2011 Games and the global polarity of the PanAmerican Approach by each country.

Fig. 5 shows the polarity of six Internet resources - Resource 1 (R1), ..., , Resource 6 (R6)- for the last thirty-three countries. For example in Fig. 5(19) the polarity of Jamaica (JAM) is positive for the first three resources, negative for the following two, and postive for the last one.

V. Conclusion and Future Work

This paper has presented an approach to analyse a subset of Internet resources focused on a specific event, the PanAmerican Games, and based on four components: a crawler, a filter, a synthesizer and a polarity analyzer.

The PanAmerican approach has the following advantages: it allows analysis of a set of real subjective expressions used by people and their polarity classification as positive, neutral, or negative. This approach reduces ambiguity and 37.5 percent of noise in the subjective elements and classifies only those, which are not identified as noise component. Also, the PanAmerican approach found out that the polarity of content producers would be influenced strongly by the results of an event with a correlation of .74. Thus, it is possible to conclude that the polarity of content producers is strongly influenced by the results of the event.

In this case, the experiments reported are of a limited scale and serve mostly to demonstrate that the PanAmerican approach is feasible. In addition, there is the potential to scale up to use with a sizeable dataset.

Furthermore, if we included all the posts of Internet resources then the PanAmerican analysis would get an accuracy less than .5. However, we developed a approach based on an polarity classification with the precision of between .8 and .9, where the precision for positive and neutral polarity are acceptable and the recall is also good. In contrast with the negative polarity where precision is higher but the recall is very low.

To conclude, one of the benefits of the results of our research is the MLO presented because it is suitable for integration with other systems.

Finally, further research should be carried out to explore geographical differences in jargon and language. For example, we aim to identify certain evaluative words that are used only in some local geographical areas.

Acknowledgements

We are grateful to the Sciences Research Council (CONACYT) for funding this research project and Multilingüismo en ontologías y linked data (BabelData), TIN2010-17550, funded by the Ministry of Science and Technology, 2011-2013.

References

[1] E. Cambria and A. Hussain, "Sentic computing: Techniques, tools, and applications," *SpringerBriefs in Cognitive Computation*, vol. 2, pp. 1–135, 2012.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found Trends Information Retrieval*, vol. 2, pp. 1–135, January 2008.

[3] S. Somasundaran, T. Wilson, J. Wiebe, and V. Stoyanov, "Qa with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news," *In International Conference on Weblogs and Social*, pp. 1–8, 2007.

[4] V. Stoyanov, C. Cardie, and J. Wiebe, "Multi-perspective question answering using the opqa corpus," *in Proceedings of HTL-EMNLP 2005*, pp. 923–930, 2005.

[5] L. V. Lita, A. H. Schlaikjer, W. Hong, and E. Nyberg, "Qualitative dimensions in question answering: Extending the definitional qa task," *In AAAI-2005*, 2005.

[6] N. Chiluka, N. Andrade, and J. Pouwelse, "A link prediction approach to recommendations in large-scale user-generated content systems," *Advances in Information Retrieval, The 33rd European Conference on Information Retrieval (ECIR 2011)*, 2011.

[7] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima, "Mining product reputations on the web," *ACM Press*, pp. 341–349, 2002.

[8] L. Ungar and D. Foster, "Clustering methods for collaborative filtering," *AAAI Press, Menlo Park California*, 1998.

[9] R. Higashinaka, M. Walker, and R. Prasad, "Learning to generate naturalistic utterances using reviews in spoken dialogue systems," *ACM Transactions on Speech and Language Processing (TSLP)*, 2007.

[10] A. Esuli and F. Sebastiani, "Sentiwordnet: A publicly available lexical resource for opinion mining," *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06*, pp. 417–422, 2006.

[11] http://nlg18.csie.ntu.edu.tw:8080/opinion/userform.jsp (Accessed: September 2013).

[12] A. Pak, "Automatic, adaptive,and applicative sentiment analysis," *Thèse de l'École Doctorale d'Informatique de l'Université Paris-Sud*, June 2012.

[13] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Opinionfinder: a system for subjectivity analysis," *Proceedings of HLT/EMNLP on Interactive Demonstrations*, pp. 34–35, 2005.

[14] A. A. Gruzd, S. Doiron, and P. Mai, "Is happiness contagious online? a case of twitter and the 2010 winter olympics." *IEEE Computer Society*, pp. 1–9, 2011.

[15] M. Thelwall, "Heart and soul: Sentiment strength detection in the social web with sentistrength," *To appear in Holyst, J. (Ed). Cyberemotions*, pp. 1–14, 2013.

[16] J. E. Campos-Quirarte, L. I. Barbosa-Santillan, and A. Castro-Munguia, "A focused crawler in order to get semantic web resources (csr)," *Thirteenth Mexican International Conference on Computer Science (ENC'13), Morelia (Mexico)*, pp. 1–6, October 2013.

[17] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine Learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[18] Y.-S. Kim, W. Street, and F. Menczer, "Feature selection in unsupervised learning via evolutionary search," *Proc. 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000.

[19] http : //www.guadalajara2011.org.mx/es/rss (Accessed: September 2013).

 ISBN: 978-1-61208-311-7

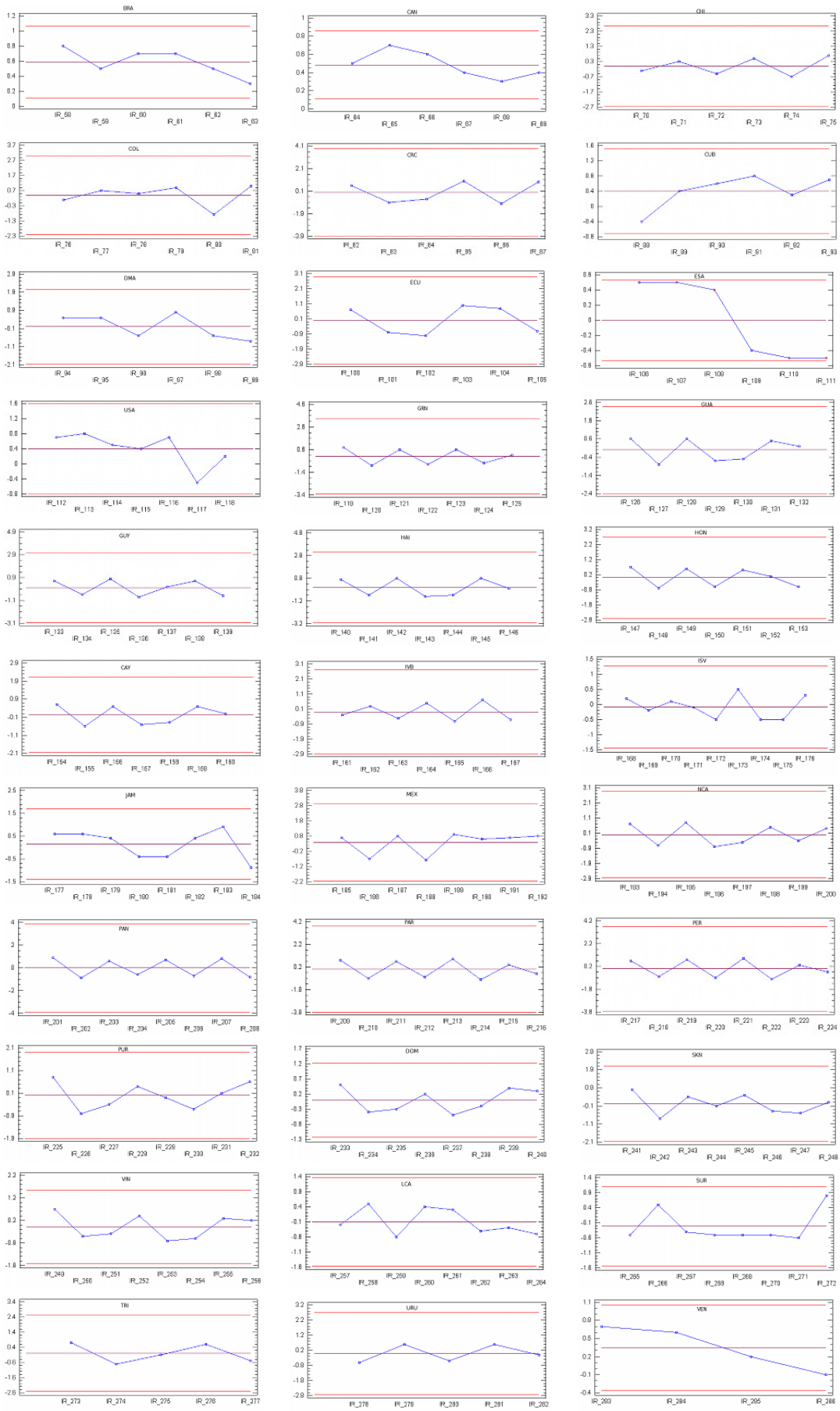

Fig. 5: Polarity of six Internet resources -Internet Resource 1 (IR1), ..., , Internet Resource 6 (IR6)- for the last thirty-three countries. For example in Fig. 5(19) the polarity of Jamaica (JAM) is positive for the first three resources, negative for the following two, and postive for the last one.